Using Decision Tree for Automatic Identification of Bengali Noun-Noun Compounds

Vivekananda Gayen[1], Kamal Sarkar[2]
[1]Department of Computer Science & Technology, Central Calcutta Polytechnic, Kolkata-700014, India
Email: vivek3gayen@gmail.com
[2]Computer Science & Engineering Department, Jadavpur University, Kolkata- 700032, India
Email: jukamal2001@yahoo.com

***Abstract*—This paper presents a supervised machine learning approach that uses a decision tree learning algorithm for recognition of Bengali noun-noun compounds as multiword expression (MWE) from Bengali corpus. Our proposed approach to MWE recognition has two steps: (1) extraction of candidate multi-word expressions using chunk information and various heuristic rules and (2) training the machine learning algorithm to recognize a candidate multi-word expression as Multi-word expression or not. A variety of association measures have been used as features for identifying MWEs. The proposed system is tested on a Bengali corpus for identifying noun-noun compound MWEs from the corpus.**

***Index Terms*—noun-noun compound, multiword expression, association measure, decision tree.**

I. Introduction

Multiword expression (MWE) from a text document can be useful for many NLP (natural language processing) applications such as information retrieval, machine translation, word sense disambiguation. Frank Samadja (1993) has defined MWEs as "recurrent combinations of words that co-occur more often than expected by chance" [1]. Timothy Baldwin et al.(2010) defined multiword expressions(MWEs) as lexical items that can be decomposed into multiple lexemes; and display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity [2]. Successful NLP applications need to identify MWEs and treat them appropriately instead of using a simple list of MWEs.

Jackendoff(1997) estimates that the number of MWEs in a native speakers's lexicon is of the same order of magnitude as the number of single words [3]. In WordNet 1.7 (Fellbaum, 1999), for example, 41% of the entries are multiword [4].

MWEs can be broadly classified into lexicalized phrases and institutionalized phrases (Ivan A. sag et al., 2002) [5]. In terms of the semantics, compositionality is an important property of MWEs. Compositionality is the degree to which the features of the parts of a MWE combine to predict the features of the whole. According to the compositionality property, the MWEs can take a variety of forms: complete compositionality (also known as institutionalized phrases, e.g. many thanks, 'রাজ্য সরকার' (Rajya Sabha, state government)), partial compositionality (e.g. light house, 'শপিং মল' (shopping mall), 'আম আদমি' (aam admi, common people)), idiosyncratically compositionality (e.g. spill the beans ,open secret, which is decomposable) and finally complete non-compositionality (e.g. hot dog, green card, 'উভয় সঙ্কট' (ubhoy sangkat, on the horns of a dilemma), which is non-decomposable).

Compound noun is a class of MWE which is rapidly expanding due to the continuous addition of new terms for introducing new ideas. Compound nouns fall into both groups: lexicalized and institutionalized. A noun-noun compound in English characteristically occurs frequently with high lexical and semantic variability (Takaaki Tanaka et al., 2003) [6]. Since compound nouns are rather productive and new compound nouns are created from day to day, it is impossible to exhaustively store all compound nouns in a dictionary.

It is also common practice in Bengali literature to use noun-noun compound as MWEs. Bengali new terms directly coined from English terms are also commonly used as MWEs in Bengali (e.g.'ডেং থ্রি' (dengue three), 'ন্যানো সিম' (nano sim), 'ভিলেজ টুযরিজম' (village tourism), 'অ্যালার্ট মেসেজ' (alert message)).

The main focus of our work is to develop a machine learning approach based on a set of statistical features for identifying Bengali noun-noun compounds.

To date, not much comprehensive work has been done on Bengali multiword expression extraction.

Previous works related to our proposed work are presented in section II. The proposed MWE identification method has been detailed in section III. The evaluation and results are presented in section IV and conclusions are drawn in last section.

II. Related Work

Multiword expression extraction can be broadly classified as: Association measure based methods, deep linguistic based methods, machine learning based methods and hybrid methods.

The earliest works on MWE extraction used statistical measures for multiword expression extraction. One of the important advantages of using statistical measures for extracting multiword expression is that these measures are language independent. Frank Smadja (1993) developed a system called Xtract that uses positional distribution and part-of-speech information of surrounding words of a word in a sentence to identify interesting word pairs [1]. Classical statistical hypothesis test like Chi-square test, t-test, z-test,

DOI: 03.LSCS.2013.3.66

log-likelihood ratio (Ted Dunning, 1993) have also been employed to extract collocations [7]. Gerlof Bouma(2009) has presented a method for collocation extraction that uses some information theory based association measures such as mutual information and point-wise mutual information [8].

Wen Zhang et al (2009) highlights the deficiencies of mutual information and suggested an enhanced mutual information based association measures to overcome the deficiencies [9]. The major deficiencies of the classical mutual information , as they mention, are its poor capacity to measure association of words with unsymmetrical co-occurrence and adjustment of threshold value. [10] Anoop et al (2008) also used various statistical measures such as point-wise mutual information(K. Church et al., 1990) [11], log-likelihood, frequency of occurrence, closed form (e.g., blackboard) count, hyphenated count (e.g., black-board) for extraction of Hindi compound noun multiword extraction. Aswhini et al (2004) has used co-occurrence and significance function to extract MWE automatically in Bengali, focusing mainly on noun-verb MWE [12]. [13] Sandipan et al (2006) has used association measures namely salience (Adam Kilgarrif et al., 2000) [14], mutual information and log likelihood for finding N-V collocation. Tanmoy (2010) has used a linear combination of some of the association measures namely co-occurrence, Phi, significance function to obtain a linear ranking function for ranking Bengali noun-noun collocation candidates and MWEness is measured by the rank score assigned by the ranking function [15].

The statistical tool (e.g., log likelihood ratio) may miss many commonly used MWEs that occur in low frequencies. To overcome this problem, some linguistic clues are also useful for multiword expression extraction. [16] Scott Songlin Paul et al (2005) focuses on a symbolic approach to multiword extraction that uses large-scale semantically classified multiword expression template database and semantic field information assigned to MWEs by the USAS semantic tagger(Paul Rayson et al.,2004) [17]. [18] R. Mahesh et al (2011) has used a stepwise methodology that exploits linguistic knowledge such as replicating words(ruk ruk e.g. stop stop), pair of words(din-raat e.g. day night), samaas(N+N,A+N) and Sandhi(joining or fusion of words), Vaalaa morpheme(jaane vaalaa e.g. about to go) constructs for mining Hindi MWEs. A Rule-Based approach for identifying only reduplication from Bengali corpus has been presented in Tanmoy et al (2010) [19]. A semantic clustering based approach for indentifying bigram noun-noun MWEs from a medium-size Bengali corpus has been presented in Tanmoy et al (2011) [20]. The authors of this paper hypothesize that the more the similarity between two components in a bigram, the less the probability to be a MWE. The similarity between two components is measured based on the overlap between the synonymous sets of the component words.

Pavel Pecina (2008) used linear logistic regression, linear discriminant analysis (LDA) and Neural Networks separately on feature vector consisting of 55 association measures for extracting MWEs [21]. M.C. Diaz-Galiano et al. (2004) has applied Kohonen's linear vector quantization (LVQ) to integrate several statistical estimators in order to recognize MWEs [22]. [23] Sriram Venkatapathy et al. (2005) has presented an approach to measure relative compositionality of Hindi noun-verb MWEs using Maximum entropy model (MaxEnt). Kishorjit et al (2011) has used a conditional random field (CRF) for Manipuri MWE extraction [24].

Hybrid methods combine statistical, linguistic and/or machine learning methods. Maynard and Ananiadou (2000) integrated both linguistics and statistical information in their system called TRUCK, for extracting multi-word terms [25]. [26] Dias (2003) has developed a hybrid system for MWE extraction, which integrates word statistics and linguistic information. [27] Carlos Ramisch et al.(2010) presents a hybrid approach to multiword expression extraction that combines the strengths of different sources of information using a machine learning algorithm. [5] Ivan A. Sag et al (2002) argued in favor of maintaining the right balance between symbolic and statistical approaches while developing a hybrid MWE extraction system.

III. Proposed Mwe Identicication Method

Our proposed MWE identification method has several major steps: preprocessing, candidate MWE extraction and MWE identification by classifying the candidates MWEs into two categories: positive (MWE) and negative (non-MWE).

A. Preprocessing

At the preprocessing step, unformatted documents are segmented into a collection of sentences automatically by checking Dari (in English, full stop), Question mark (?) and Exclamation sign (!). Typographic or phonetic errors are not corrected automatically. Then the sentences are submitted to the chunker [1] one by one for processing.

B. Candidate MWE Extraction

The chunked sentences are processed to identify the multi-word expression candidates. The multiword expression candidates are primarily extracted using the following rule:

Bigram consecutive token sequence within same NP chunk is extracted from the chunked sentences if the Tag of the token is NN or NNP or XC (NN: Noun, NNP: Proper Noun, XC: compounds (Akshar Bharati et al., 2006)) [28].

We observed that some potential multi- word expressions are missed due to the chunker's error. For example, the chunked version of the sentence is ((NP কবেকার NN)) ((NP বিএসএ NN)) ((NP সাইকেল NN , SYM)). In this example, we find that the potential multi-word expression candidate "বিএসএ সাইকেল" (BSA Cycle) cannot be detected using the above rule since "বিএসএ" (BSA) and "সাইকেল" (Cycle) belong to the different chunks.

To identify more number of potential MWE candidates, we use some heuristic rules as follows:

Bigram noun-noun compounds which are hyphenated or occur within single quote or within first brackets or whose words are out of vocabulary (OOV) are also considered as

DOI: 03.LSCS.2013.3.66

the potential candidates for MWE.

C. Features

The association measures namely phi, point-wise mutual information (pmi), salience, log likelihood, poisson stirling, chi and t-score have been used to calculate the scores of each candidate MWE. These association measures use various types of frequency statistics associated with the bigram. The frequency statistics used in computing association measures are represented using a typical contingency table format (Satanjeev Banerjee et al., 2003) [29]. Table 1 shows a typical contingency table showing various types of frequencies associated with the bigram <word1, word2> (e.g., রাজ্য সরকার).

TABLE I. CONTINGENCY TABLE

	সরকার (government)	~সরকার (~ government)	
রাজ্য (state)	n_{11}	n_{12}	n1p
~রাজ্য (~state)	n_{21}	n_{22}	n2p
	np1	np2	npp

The meanings of the entries in the contingency table are given below:

n_{11} = number of times the bigram occurs, joint frequency.

n_{12} = number of times word1 occurs in the first position of a bigram when word2 does not occur in the second position.

n_{21} = number of times word2 occurs in the second position of a bigram when word1 does not occur in the first position.

n_{22} = number of bigrams where word1 is not in the first position and word2 is not in the second position.

n1p = the number of bigrams where the first word is word1, that is, n1p = n_{11} + n_{12}.

np1 = the number of bigrams where the second word is word2, that is np1=n_{11}+n_{21}.

n2p = the number of bigrams where the first word is not word1, that is n2p=n_{21}+n_{22}.

np2 = the number of bigrams where the second word is not word2, that is np2=n_{12}+n_{22}.

npp is the total number of bigram in the entire corpus.

Using the frequency statistics given in the contingency table, expected frequencies, m_{11}, m_{12}, m_{21} and m_{22} are calculated as follows:

$$m_{11} = (n1p*np1/npp)$$

$$m_{12} = (n1p*np2/npp)$$

$$m_{21} = (np1*n2p/npp)$$

$$m_{22} = (n2p*np2/npp)$$

where:

m_{11}: Expected number of times both words in the bigram occur together if they are independent.

m_{12}: Expected number of times word1 in the bigram will occur in the first position when word2 does not occur in the second position given that the words are independent.

m_{21}: Expected number of times word2 in the bigram will occur in the second position when word1 does not occur in the first position given that the words are independent.

m_{22}: Expected number of times word1 will not occur in the first position and word2 will not occur in the second position given that the words are independent.

The following association measures that use the above mentioned frequency statistics are used in our experiment.

Phi, Chi and T-score

The Phi, Chi and T-score are calculated using the following equations:

$$phi = \frac{((n_{11}*n_{22})-(n_{12}*n_{21}))}{\sqrt{(n1p*np1*np2*n2p)}} \quad (1)$$

$$chi = 2*\left(\left(\frac{(n_{11}-m_{11})}{m_{11}}\right)^2+\left(\frac{(n_{12}-m_{12})}{m_{12}}\right)^2+\left(\frac{(n_{21}-m_{21})}{m_{21}}\right)^2+\left(\frac{(n_{22}-m_{22})}{m_{22}}\right)^2\right) \quad (2)$$

$$T-Score = \frac{(n_{11}-m_{11})}{\sqrt{n_{11}}} \quad (3)$$

Log likelihood, Pmi, Salience and Poisson Stirling

Log likelihood is calculated as:

$$LL = 2*(n_{11}*\log(n_{11}*m_{11})+n_{12}*\log(n_{12}*m_{12})+n_{21}*\log(n_{21}*m_{21})+n_{22}*\log(n_{22}*m_{22})) \quad (4)$$

Point-wise Mutual Information (pmi) is calculated as:

$$pmi = \log\left(\frac{n_{11}}{m_{11}}\right) \quad (5)$$

The salience is defined as:

$$salience = \left(\log\left(\frac{n_{11}}{m_{11}}\right)\right) * \log(n_{11}) \quad (6)$$

The Poisson Stirling measure is calculated using the formula:

$$Poisson-Stirling = n_{11} * \left(\left(\log\left(\frac{n_{11}}{m_{11}}\right)\right)-1\right) \quad (7)$$

Co-occurrence

Co-occurrence is calculated using the following formula (Aswhini Agarwal et al., 2004) [12]:

$$co(w1,w2) = \sum_{s \in S(w1,w2)} e^{-d(s,w1,w2)} \quad (8)$$

Where:

co(w1,w2) = co-occurrence between the words (after stemming).

S(w1,w2)= set of all sentences where both w1 and w2 occurs.

d(s,w1,w2)= distance between w1 and w2 in a sentence in terms of words.

Significance Function

The significance function (Aswhini Agarwal et al., 2004) [12] is defined as:

DOI: 03.LSCS.2013.3.66

$$sig_{w1}(w2) = \sigma[k1(1 - co(w1, w2).\frac{f_{w1}(w2)}{f(w1)})].\sigma[k2.\frac{f_{w1}(w2)}{\lambda} - 1] \quad (9)$$

$$sig(w1, w2) = sig_{w1}(w2).\exp[\frac{f_{w1}(w2)}{\max(f_{w1}(w2))} - 1] \quad (10)$$

Where:

$sig_{w1}^{(w2)}$ = significance of w2 with respect to w1.

$f_{w1}^{(w2)}$ = number of w1 with which w2 has occurred.

Sig(w1,w2)= general significance of w1 and w2 , lies between 0 and 1.

σ(x)= sigmoid function =exp(-x)/(1+exp(-x))]

k1 and k2 define the stiffness of the sigmoid curve (for simplicity they are set to 5.0)

λ is defined as the average number of noun-noun co-occurrences.

D. MWE Identification using a Decision Tree

Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented in the form of a decision tree. Decision trees are supervised algorithms which recursively partition the data, based on its attributes, until some stopping condition is reached. This recursive partitioning gives rise to a tree-like structure. A decision tree is a tree where the non-leaf nodes are labeled with attributes. The arcs from a node representing the attribute *A*, are labeled with each of the possible values of the attribute *A*. The leaves of the tree are labeled with classifications. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance (Mitchell, 1997) [30]. An instance is initially classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute.

The most important feature of a decision tree classifier is its capability to break down a complex decision–making process into a collection of simpler decisions, thus providing a solution, which is often easier to interpret (Safavian et al., 1991) [31].

C4.5 is an algorithm developed by Ross Quinlan(Quinlan,1993) [32]. This is used to generate a decision tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm (Quinlan,1986) [33]. For our MWE identification task, the C4.5 decision tree is trained to classify the candidate MWEs in a document as one of two categories: "MWE", "not a MWE".

Training a decision tree learning algorithm for MWE identification requires candidate MWEs to be represented as the feature vectors. For this purpose, we write a computer program for automatically extracting values for the features characterizing the MWE candidates in the documents. For each candidate MWE in a document in our corpus, we extract the values of the features of the candidate using the measures discussed in subsection C of section III. If the candidate MWE is found in the list of manually identified MWEs, we label the MWE as a "Positive" example and if it is not found we label it as a "negative" example. Thus the feature vector for each candidate looks like {<a_1 a_2 a_3 ….. a_n>, <label>} which becomes a training instance (example) for the decision tree, where a_1, a_2 . . .a_n, indicate feature values for a candidate. A training set consisting of a set of instances of the above form is built up by running a computer program on the documents in our corpus.

For our experiment, we use Weka (www.cs.waikato.ac.nz/ml/weka) machine learning tools. We use J48, which is C4.5 version of the decision tree under WEKA workbench, included under the panel Classifier/ trees of WEKA workbench. For our work, the J48 classifier of the WEKA suite has been run with the default values of its parameters.

IV. Evaluation And Results

For evaluating the performance of our system the traditional precision, recall and F-measure are computed by comparing machine assigned labels to the human assigned labels for the candidate MWEs extracted from our corpus of 274 Bengali documents.

A. Experimental dataset

Our corpus is created by collecting the news articles from the online version of well known Bengali newspaper ANANDABAZAR PATRIKA during the period spanning from 20.09.2012 to 19.10.2012. The news articles published online under the section Rajya (State) , Desh on the topics bandh-dharoghat, crime, disaster, jongi, mishap, political and miscellaneous are included in the corpus. It consists of total 274 documents and all those documents contain 18769 lines of Unicode texts, 233430 tokens. We have manually identified all the noun-noun compound MWEs in the collection and created a gold standard for labeling the training data. It consists of 4641 noun-noun compound MWEs. Total 8210 noun-noun compound MWE candidates are automatically extracted employing chunker and heuristic rules as described in subsection B of section III.

B. Results

To estimate overall accuracy of our proposed MWE identification system, 10-fold cross validation is done. The dataset is randomly reordered and then split into *n* parts of equal size. For each of 10 iterations, one part is used for testing and the other *n*-1 parts are used for training the classifier. The test results are collected and averaged over all folds. This gives the cross-validation estimate of the accuracy of the proposed system. J48 which is basically a decision tree included in WEKA is used as a single decision tree for implementing our system. Our proposed decision tree based system gives an average F-measure of 0.77.

[1]http//ltrc.iiit.ac.in/analyzer/bengali

DOI: 03.LSCS.2013.3.66

Conclusions And Future Work

This paper presents a machine learning based approach for identifying noun-noun compound MWEs from a Bengali corpus. We have used a number of association measures as features which are combined by a decision learning algorithm for recognizing noun-noun compounds.

As a future work, we have planned to improve the candidate MWE extraction step of the proposed system and/ or introduce new features such as lexical features and semantic features.

References

[1] Frank Smadja 1998. "Retrieving Collocation from Text: Xtract." *Computational Linguistics*. 19.1(1993):143-177.

[2] Timothy Baldwin and Su Nam Kim (2010), in Nitin Indurkhya and Fred J. Damerau (eds .) Handbook of Natural Language Processing, *Second Edition, CRC Press*, Boca Raton, USA, pp. 267-292 .

[3] Jackendoff, Ray: 1997, The Architecture of the Language Faculty, *Cambridge, MA: MIT Press*.

[4] Fellbaum, Christine, ed.: 1998, WordNet: An Electronic Lexical Database, *Cambridge,MA: MIT Press*.

[5] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger.2002. Multi-worh expression: A Pain in the neck for NLP. *CICLing*, 2002.

[6] Takaaki Tanaka, Timothy Baldwin. 2003. "Noun-Noun Compound Machine Translation: a Feasibility Study on Shallow Processing." *Proceeings of the ACL 2003 workshop on Multiword expressions*. pp. 17-24.

[7] Ted Dunning. 1993. Accurate Method for the Statistic of Surprise and Coincidence. *In Computational Linguistics*, pp. 61-74.

[8] Gerlof Bouma. 2009. "Normalized (pointwise) mutual information in collocation extraction." *Proceedings of GSCL* (2009): 31-40.

[9] Wen Zhang, Taketoshi Yoshida, Xijin Tang, Tu-Bao Ho. 2009. Improving effictiveness of mutual information for substantival multiword expression extraction. Expert Systems with Applications 36 (2009) 10919-10930 (ELSEVIER).

[10] Anoop Kunchukuttan and Om P. Damani. 2008. A System for Compound Noun Multiword Expression Extraction for Hindi. *In proceeding of 6th International Conference on National Language Processing (ICON)*. pp. 20-29.

[11] K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*. 16(1). 1990.

[12] Aswhini Agarwal, Biswajit Ray, Monojit Choudhury, Sudeshna Sarkar and Anupam Basu. 2004. Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenario. *In Proceedings of International Conference on Natural Language Processing (ICON)*, pp. 165-17.

[13] Sandipan Dandapat, Pabitra Mitra and Sudeshna Sarkar. 2006. Statistical Investigation of Bengali Noun-Verb (N-V) Collocations as Multi-word expressions. *In the Proceedings of MSPIL*, Mumbai, pp 230-233.

[14] Adam Kilgarrif and Joseph Rosenzweig. 2000. Framework and Results for English Senseval. *Computer and the Humanities*, 34(1): pp 15-48.

[15] Tanmoy Chakraborty. 2010. Identification of Noun-Noun(N-N) Collocations as Multi-Word Expressions in Bengali Corpus. *8th International Conference on Natural Language Processing (ICON 2010)*.

[16] Scott Songlin Piao, Paul Rayson, Dawn Archer, Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language (ELSEVIER) 19 (2005)* pp. 378-397.

[17] Paul Rayson, Dawn Archer, Scott Piao and Tony McEnery. 2004. The UCREL semantic analysis system. *In Proceedings of the LREC-04 Workshop, beyond Named Entity Recognition Semantic Labelling for NLP Tasks*, Lisbon, Portugal, pp.7-12.

[18] R. Mahesh and K. Sinha. 2011. Stepwise Mining of Multi-Word Expressions in Hindi. *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)* pp. 110-115.

[19] Tanmoy Chakraborty and Sivaji Bandyopadhyay. 2010. Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule-Based Approach. *Proceedings of Workshop on Multiword Expressions: from Theory to Applications* (MWE 2010) pp. 72-75.

[20] Tanmoy Chakraborty, Dipankar Das and Sivaji Bandyopadhyay. 2011. Semantic Clustering: an Attempt to Identify Multiword Expressions in Bengali. Proceedings of Workshop on Multiword Expressions: from Parsing and Generation to the Real World(MWE 2011*). Association for Computational Linguistics*. Portland, Oregon, USA, 23 June 2011.

[21] Pavel Pecina. 2008. Reference data for czech collocation extraction. *In Proc. Of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008*). pp. 11-14, Marrakech, Morocco, Jun.

[22] M.C. Diaz-Galiano, M.T. Martin-Valdivia, F. Martinez-Santiago, L.A. Urea-Lopez. 2004. Multiword Expressions Recognition with the LVQ Algorithm. *Workshop on methodologies and evaluation of Multiword Units in Real-word Applications associated with the 4th International Conference on Languages Resources and Evaluation*, Lisbon, Portugal. pp.12-17.

[23] Sriram Venkatapathy, Preeti Agrawal and Aravind K. Joshi. Relative Compositionality of Noun+Verb Multi-word Expressions in Hindi. *In Proceedings of ICON-2005*, Kanpur.

[24] Kishorjit Nongmeikapam, Ningombam Herojit Singh, Bishworjit Salam and Sivaji Bandyopadhyay. 2011. Transliteration of CRF Based Multiword Expression (MWE) in Manipuri: From Bengali Script Manipuri to Meitei Mayek (Script) Manipuri. *International Journal of Computer Science and Information Technology*, vol.2(4) . pp. 1441-1447.

[25] Maynard D., Ananiadou S. 2000. Trucks: a model for automatic multiword term recognition. Journal of Natural Language Processing 8 (1), 101-126.

[26] Dias D. 2003. Multiword unit hybrid extraction. In Proceedings of the Workshop on Multiword Expressions: Aanlysis, Acquisition and Treatment, at ACL 2003, Sapporo, Japan, pp. 41-48.

[27] Carlos Ramisch, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, Maria José Finatto: *A* Hybrid Approach for Multiword Expression Identification. *PROPOR 2010*: 65-74.

DOI: 03.LSCS.2013.3.66

[28] Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal. 2006. AnnCorra : Annotating Corpora Guidelines For POS And Chunk Annotation For Indian Languages.

[29] Santanjeev Banerjee and Ted Pedersen. 2003. "The Design, Implementation and Use of the Ngram Staistics Package." *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics.* Pp. 370-381.

[30] Mitchell T. M., Machine Learning, McGraw-Hill International Editions,1997.

[31] Safavian S. R. and Landgrebe D., "A Survey of Decision Tree Classifier Methodology", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 21, issue 3, May/June, , pp. 660-674, 1991.

[32] Quinlan J. R., "C4.5: Programs for Machine Learning", *Morgan Kaufmann Publishers*, 1993.

[33] Quinlan J. R., "Induction of decision trees", *Machine Learning*, 1(1), 81 – 106, 1986.

DOI: 03.LSCS.2013.3.66